گزارش پروژه‌ی ماشین لرنینگ

آرمان چم‌حیدری (۹۹۲۴۳۰۳۰) - سنا حقیقی (۹۹۲۴۳۰۸۲)

زمستان ۱۴۰۲

مرحله اول در پیش‌پردازش داده‌ها، جمع‌آوری و تهیه داده‌های مورد نیاز برای آموزش مدل است. این مرحله شامل فعالیت‌هایی مانند جمع‌آوری داده‌ها، تمیز کردن داده‌ها، تقسیم داده‌ها به دسته‌های آموزش و ارزیابی و غیره است.

- جمع‌آوری داده‌ها: در این مرحله، داده‌های مورد نیاز برای آموزش مدل جمع‌آوری می‌شوند. این داده‌ها می‌توانند شامل متن، تصاویر، صدا یا دیگر انواع داده‌ها باشند. داده‌ها می‌توانند از منابع مختلف مانند پایگاه‌های داده، وبسایت‌ها، فایل‌ها و سایر منابع باشند.
- تمیز کردن داده‌ها: در این مرحله، داده‌ها مورد تمیز کردن قرار می‌گیرند. ممکن است داده‌ها دارای نویز، اطلاعات نامربوط، مقادیر نامناسب و غیره باشند که باید از مجموعه داده حذف شوند. به عنوان مثال، در متنوع‌سازی نوع، نشانه‌های نگارشی غیرضروری مانند علامت‌های نقطه‌گذاری و ویرگول‌ها حذف می‌شوند. همچنین، داده‌های تکراری نیز می‌توانند حذف شوند تا از تکرار اطلاعات جلوگیری شود.
- تقسیم داده‌ها: داده‌ها به طور معمول به دو بخش تقسیم می‌شوند: مجموعه داده آموزش و مجموعه داده ارزیابی (یا تست). مجموعه داده آموزش برای آموزش مدل استفاده می‌شود و مجموعه داده ارزیابی برای ارزیابی عملکرد مدل در داده‌هایی استفاده می‌شود که در مرحله آموزش استفاده نشده‌اند. تقسیم داده‌ها به این دو بخش به منظور ارزیابی صحیح عملکرد مدل و اندازه‌گیری دقت آن در داده‌های جدید است.
- تعادل داده‌ها (در صورت لزوم): در برخی موارد، داده‌ها ممکن است ناهمتوازن باشند، به این معنی که تعداد نمونه‌های یک دسته نسبت به سایر دسته‌ها بسیار کمتر یا بیشتر باشد. در این صورت، معمولاً اقداماتی مانند افدموسازی داده‌ها (data augmentation) وجود دارد تا تعادل بین دسته‌ها برقرار شود. دموسازی داده‌ها شامل تغییراتی است که به نمونه‌های آموزشی اعمال می‌شود، مانند چرخش تصویر، تغییرات اندازه، تغییر رنگ و غیره. این تغییرات باعث افزایش تنوع داده‌ها و ایجاد نمونه‌های جدید می‌شود.
- نرمال‌سازی داده‌ها: در این مرحله، داده‌ها مورد پیش‌پردازش قرار می‌گیرند تا در محدوده‌ای مشخص قرار بگیرند. این مرحله به منظور کاهش تفاوت‌های موجود در داده‌ها و راحت‌تر کردن آموزش مدل استفاده می‌شود. به عنوان مثال، مقادیر عددی می‌توانند نرمال‌سازی شوند با تقسیم بر میانگین و انحراف معیار آنها.

- با انجام این مراحل، داده‌ها آماده شده و می‌توانند برای آموزش مدل استفاده شوند. البته، مراحل پیش‌پردازش داده ممکن است بسته به مسئله و نوع داده‌ها متفاوت باشند و برای هر مسئله ممکن است نیاز به مراحل دیگری نیز وجود داشته باشد.

مرحله دوم:

در این مرحله، ما باید عملیات پیش پردازش را انجام دهیم. پیش پردازش مجموعه داده به معنای اعمال تغییرات و تبدیلاتی بر روی داده‌ها است تا بهترین شرایط را برای آموزش مدل فراهم کنیم. به طور کلی، می‌توانیم این مرحله را به چندین بخش تقسیم کنیم:

تبدیل تصاویر به فرمت مناسب: تصاویر را باید به فرمتی که مدل قابل فهم است، تبدیل کنیم. معمولاً تصاویر را به آرایه‌های ۲ بعدی از اعداد در بازه [۰، ۱] تبدیل می‌کنیم. می‌توانید از کتابخانه‌های مورد علاقه خود برای این تبدیل استفاده کنید.

نرمال سازی داده‌ها: برای استفاده بهینه از شبکه عصبی، معمولاً داده‌ها را نرمال سازی می‌کنیم. این به معنای تغییر مقیاس داده‌ها به یک بازه مشخص است که معمولاً [-۱، ۱] یا [۰، ۱] است.

تقسیم مجموعه داده: قبل از آموزش مدل، باید مجموعه داده را به دو بخش آموزش و آزمون تقسیم کنیم. معمولاً یک قسمت از داده‌ها را برای آموزش مدل استفاده می‌کنیم و قسمت دیگر را برای ارزیابی عملکرد مدل نگه می‌داریم. این کار به ما کمک می‌کند تا میزان دقت و عملکرد مدل را در زمان اجرا بررسی کنیم.

مرحله دوم در پیش‌پردازش داده‌ها، تمیز کردن و تبدیل کردن داده‌ها به فرمت مناسب برای استفاده در مدل است. این مرحله شامل انجام عملیاتی مانند پاکسازی داده، نرمال‌سازی، تبدیل داده‌ها به بردارهای عددی و غیره است.
پاکسازی داده شامل حذف داده‌های مفقود، تکراری یا نامناسب است. به عنوان مثال، ممکن است بخشی از داده‌ها مقادیر نامعتبر داشته باشد یا برخی از داده‌ها تکراری باشند که می‌بایست از مجموعه داده حذف شوند.

نرمال‌سازی داده به معنای تغییر مقیاس داده‌ها به گونه‌ای است که مقادیر آنها در یک بازه مشخص قرار گیرند. این کار معمولاً با استفاده از روش‌هایی مانند مقیاس‌بندی مین-مکس یا استانداردسازی صورت می‌گیرد. این کار می‌تواند بهبودی در عملکرد مدل بیاورد، زیرا می‌تواند مشکل تغییر مقیاس متغیرهای ورودی را حل کند و به مدل کمک کند تا بهتری درک از داده‌ها و الگوهای موجود داشته باشد.
تبدیل داده‌ها به فرمت مناسب برای استفاده در مدل، به عنوان مثال تبدیل متن به بردارهای عددی با استفاده از روش‌هایی مانند کدگذاری واژگانی (مانند تبدیل واژه به بردارهای جاسازی) یا تبدیل تصاویر به بردارهای ویژگی (مانند استخراج ویژگی‌ها با استفاده از شبکه‌های عصبی عمیق) است.
اگر از data augmentation استفاده نشود، برخی از اتفاقات ممکن است رخ دهد:

کمبود تنوع: بدون استفاده از data augmentation، ممکن است داده‌ها بسیار محدود و تکراری باشند. این موضوع می‌تواند منجر به کمبود تنوع در داده‌ها شود و مدل را با الگوهای عمومی تری آشنا نکند. این ممکن است باعث کاهش قدرت تعمیم مدل شود و در نتیجه عملکرد آن بر روی داده‌های جدید ناپایدار شود.

بیش‌برازش (overfitting): وقتی که داده‌های آموزش کم و تکراری هستند، مدل می‌تواند به طور غیرمنطقی و بیش از حد به اعملکرد آنها با داده‌های آموزش سازگار شود. این موضوع ممکن است باعث بیش‌برازش شود، به این معنی که مدل به درستی روی داده‌های آموزش عمل کند، اما نتواند به درستی روی داده‌های جدید و نامعتبر پاسخگو باشد.

ناپایداری: بدون استفاده از data augmentation، ممکن است مدل در مواجهه با تغییرات و تنوع‌های مختلف در داده‌های جدید ناپایدار عمل کند. این می‌تواند منجر به بالا بردن خطاها و کاهش دقت مدل شود.

به طور کلی، استفاده از data augmentation به ما کمک می‌کند تا تنوع بیشتری در داده‌ها ایجاد کنیم و مدل را در برابر تغییرات و تنوع‌های مختلف مقاوم کنیم. این تکنیک معمولاً بهبودی در عملکرد مدل می‌آورد و به ما اجازه می‌دهد تا با دقت بیشتری پیش‌بینی کنیم. بدون استفاده از این تکنیک، ممکن است مدل در مواجهه با داده‌های جدید ناپایدار عمل کند و به نتایج ناقصی برسد.

مرحله سوم:
در این مرحله، ما باید یک شبکه عصبی مصنوعی (ANN) را پیاده سازی کنیم تا بتواند ارقام فارسی را با توجه به مجموعه داده جمع آوری شده تشخیص دهد. شبکه عصبی مصنوعی یک مدل ریاضی است که توسط لایه‌هایی از نورون‌ها ساخته می‌شود. هر نورون وزن‌هایی دارد که مقدار ورودی را تغییر می‌دهد و به اعداد خروجی مربوطه می‌رساند. این وزن‌ها در فرآیند آموزش مدل توسط الگوریتم‌های بهینه‌سازی بهبود می‌یابند.
مرحله سوم در پیش‌پردازش داده‌ها، تقسیم مجموعه داده به بخش‌های آموزش و آزمون است. این کار به ما کمک می‌کند تا مدل را بر روی داده‌های آموزش آموزش دهیم و سپس عملکرد آن را با استفاده از داده‌های آزمون ارزیابی کنیم.
تقسیم داده به دو بخش اصلی می‌تواند به صورت تصادفی انجام شود، به طوری که مثلاً 80٪ از داده‌ها را برای آموزش استفاده کنیم و 20٪ را برای آزمون نگه داریم. این نسبت می‌تواند بر اساس محدودیت‌های مسئله و اندازه مجموعه داده تعیین شود.
دلیل استفاده از تقسیم داده به بخش‌های آموزش و آزمون این است که بتوانیم عملکرد مدل را بر روی داده‌هایی که قبلاً دیده نشده است، ارزیابی کنیم. این کار به ما اطمینان می‌دهد که مدل به طور عمومی عمل می‌کند و قابلیت تعمیم را دارد.
درباره data augmentation، این تکنیک معمولاً در مسائل بینایی ماشین استفاده می‌شود و به ما کمک می‌کند تا مجموعه داده را با اعمال تغییرات کوچک به تصاویر، گسترش دهیم. به عنوان مثال، می‌توان تصاویر را به صورت افقی و عمودی برگرداند، آنها را با زوایای مختلف چرخاند، اندازه‌های مختلفی برای تصویر انتخاب کنیم، و یا روش‌های دیگری مانند نویز اضافه کردن، برش دادن و جابجایی کردن اعمال کنیم.
استفاده از data augmentation می‌تواند به ما کمک کند تا تنوع بیشتری در داده‌ها ایجاد کنیم و مدل را در برابر تغییرات و تنوع‌های مختلف مقاوم کنیم. این تکنیک معمولاً

بهبودی در عملکرد مدل می‌آورد، زیرا با افزایش تنوع داده‌ها، مدل قادر است الگوهای عمومی تری را درک کند و بهتر در مقابل داده‌های جدید عمل کند.
بدون استفاده از data augmentation، ممکن است مدل ما به علت کمبود تنوع در داده‌ها، به عملکرد نسبتاً ضعیفی برخورد کند. ممکن است مدل عمومی‌تای را نیاز به یک تنوع بیشتر در داده‌ها داشته باشد و نتواند الگوهای عمومی را به خوبی یاد بگیرد. این می‌تواند منجر به بالا بردن خطاها و کاهش دقت مدل در مواجهه با داده‌های جدید شود.
به طور کلی، استفاده از data augmentation بهبودی در عملکرد مدل می‌آورد و به ما اجازه می‌دهد تا با دقت بیشتری پیش‌بینی کنیم. بدون استفاده از این تکنیک، ممکن است مدل ما در مواجهه با داده‌های جدید ناپایدار عمل کند و به نتایج ناقصی برسد.

مرحله چهارم: طراحی شبکه عصبی پیچشی (CNN) برای تشخیص ارقام فارسی

ورودی شبکه: برای طراحی شبکه عصبی پیچشی برای تشخیص ارقام فارسی، ابتدا باید ورودی مناسب را تعیین کنیم. در اینجا، می‌توانیم تصاویر ارقام فارسی را به عنوان ورودی استفاده کنیم. اندازه تصاویر و نوع فرمت آنها باید با مجموعه داده ما همخوانی داشته باشد.

لایه‌های پیچشی: شبکه عصبی پیچشی تشکیل شده از لایه‌های پیچشی است که به واسطه فیلترها (kernel) و عملیات پیچش، اطلاعات محلی را از تصاویر استخراج می‌کنند. هر لایه پیچشی شامل یک تعدادی فیلتر است که روی تصویر عمل پیچش انجام می‌دهند و نتایج را به لایه بعد منتقل می‌کنند. معمولاً این لایه‌ها با استفاده از تابع فعال‌سازی ReLU (Rectified Linear Unit) به‌کار می‌روند.

لایه‌های تجمیع (Pooling): لایه‌های تجمیع برای کاهش ابعاد فضایی تصاویر استفاده می‌شوند و اطلاعات مهم را از تصاویر استخراج می‌کنند. این لایه‌ها با استفاده از روش‌هایی مانند Max Pooling یا Average Pooling، بخش‌هایی از تصویر را که اطلاعات کمتری دارند را حذف کرده و اطلاعات مهم را حفظ می‌کنند.

لایه‌های کاملا متصل: پس از لایه‌های پیچشی و تجمیع، اطلاعات استخراج شده به لایه‌های کاملا متصل (Fully Connected) منتقل می‌شوند. این لایه‌ها اطلاعات را ترکیب و به یک لایه خروجی می‌فرستند که به ما کمک می‌کند ارقام فارسی را

تشخیص دهیم. معمولاً این لایه خروجی از تابع فعال‌سازی softmax برای تولید احتمالات مربوط به هر ارقام استفاده می‌کند.

آموزش مدل: بعد از طراحی شبکه عصبی، باید مدل را با استفاده از مجموعه داده آموزش آموزش دهیم. در این مرحله، مجموعه داده آموزش به مدل داده می‌شود و شبکه عصبی با بهره‌گیری از الگوریتم بهینه‌سازی، ماز مجموعه داده آموزش، وزن‌های شبکه را به‌روزرسانی می‌کند تا بتواند ارقام فارسی را تشخیص دهد.

مرحله پنجم: افزایش داده (Data Augmentation) و ارزیابی مدل

جستجوی افزایش داده: در این مرحله، شما باید روش‌های افزایش داده را برای مجموعه داده خود جستجو کنید. افزایش داده به روش‌هایی مانند چرخش، تغییر اندازه، تغییر روشنایی و غیره اعمال می‌شود. این روش‌ها به شما کمک می‌کنند تا تنوع بیشتری در داده‌های آموزش خود ایجاد کنید و از برازش زیاد (overfitting) جلوگیری کنید.

اعمال افزایش داده: پس از جستجو و انتخاب روش‌های افزایش داده، باید آنها را بر روی مجموعه داده خود اعمال کنید. به طور معمول، این کار با استفاده از کتابخانه‌های پردازش تصویر مانند OpenCV یا TensorFlow انجام می‌شود. با اعمال این روش‌ها، تعداد داده‌های آموزش افزایش می‌یابد و مدل شما قادر خواهد بود الگوهای بیشتری را یاد بگیرد.

ارزیابی مدل: پس از اعمال افزایش داده، باید مدل را دوباره ارزیابی کنید. برای این کار، می‌توانید مجموعه داده ارزیابی خود را به مدل بدهید و عملکرد آن را اندازه‌گیری کنید. می‌توانید از معیارهایی مانند دقت (accuracy)، ماتریس درهم‌ریختگی (confusion matrix) و منحنی مشخصه مشترک (ROC curve) استفاده کنید.

مقایسه نتایج: در این مرحله، باید نتایج مدل پس از اعمال افزایش داده را با نتایج قبلی مقایسه کنید. اگر اعمال افزایش داده باعث بهبود عملکرد مدل شده است، شما می‌توانید از مدل با افزایش داده آموزش دیده برای تشخیص ارقام فارسی استفاده کنید.

اهمیت افزایش داده در آموزش مدل‌های عصبی این است که با افزایش تنوع داده‌ها، مدل قادر خواهد بود الگوهای مختلف را یاد بگیرد و بهتر در مقابل داده‌های جدید و ناشناخته عمل کند.